Toward a Science of Robotics: Goals and Standards for Experimental Research

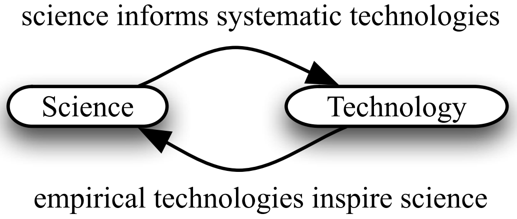
Leila Takayama
Willow Garage
68 Willow Road, Menlo Park, CA 94025
Email: takayama@willowgarage.com

***Abstract*— Drawing from lessons learned in the combination of sciences and engineering in the field of human-computer interaction, this paper presents a set of issues relating to improving the scientific rigor of experiments in robotics, moving toward a sciences of robotics. It highlights the strengths of a variable-based approach to the study of technologies in comparison to A/B testing. Two specific examples of human-robot interaction experiments are presented in terms of the methodological design decisions made in the process of conducting that published research. Finally, a broader discussion of experimental methodologies addresses issues of experiment designs, statistical analyses, reporting methods and results, and other experiment work practices that may provide guidelines and work practices from other disciplines to improve the standards of experimental research in robotics.**

I. Introduction

The fusion of science and technology is a challenge faced by many engineering research efforts, including robotics. One particularly useful frame for thinking about the relationships between science and technology was articulated by Michael Polanyi, a philosopher who coined the terms "empirical technology" and "systematic technology" [12]. Empirical technologies are those technologies that are largely unscientific, but are so unique and innovative that they inspire new scientific theories and explorations, thereby contributing to science. Systematic technologies are those that are deeply informed by current scientific knowledge, thereby benefiting from science. In an ideal world, any given robotic system would contain the best of both–incorporating the state of the art in scientific knowledge and inspiring fundamentally new research questions. This balance of science and technology facilitates a feedback cycle between the engineering work of inventing new technologies, and the scientific work of generating new knowledge.

Fig. 1. Polanyi's Relationships Between Science and Technology


science informs systematic technologies
Science
Technology
empirical technologies inspire science

One of the challenges of doing rigorous scientific research in robotics is creating a body of knowledge through proper scientific experimentation. As thoroughly discussed by Simon [15], though the sciences of the natural world are often held up as gold standards (e.g., physics), there are also rigorous ways to study artificial things (i.e., things that are synthesized by human beings, p. 5), which may be different. This suggests that looking at other sciences of the artificial, e.g., human-computer interaction (HCI) and human-robot interaction (HRI), might be helpful for informing good experimental research practices in the broader robotics community. This paper aims to take a concrete step toward articulating the goals and standards necessary for meeting the challenge of building a rigorous science of robotics from an HCI/HRI perspective.

II. Goals and Standards

One goal may be to build robotic systems that are both systematic and empirical technologies [12] in order to develop a science of robotics. In other words, these robotic systems should be deeply informed by existing scientific knowledge, and also inspire new directions of scientific inquiry. To this end, it is informative to review the standards by which scientific experiments are designed, conducted, and presented.

Among the many standards of scientific research is the expectation that phenomena are observable, repeatable, and objective (or, at least, free of as many sources of bias as possible). The implications of these standards are that experiments must:

1) use good measures, which are standardized and/or usable by others, and are as objective as possible
2) clearly and thoroughly explain the exact methods used to run the study so that others can reproduce the study to see if they get similar or different results
3) minimize subjectivity in experimental observations

Many of the lessons learned from methods courses in experimental social science research are directly applicable to the methods that could improve the quality of experimental research in the robotics community. For example, creating standardized and thoroughly tested sets of measures (e.g., handbooks of research measures) are common practice in fields such as Cognitive Psychology, Social Psychology, and Communication. These measures are tested for many types of validity, including:

- **face validity**: how reasonable a measure seems to be for its concept
- **content validity**: how thoroughly a measure addresses the breadth of a concept
- **construct validity**: how much a measure causally relates to other variables within one's theory
- **external validity**: how generalizable the results will be to other systems and contexts

All of these validity criteria involve how valid and generalizable one's research results may be. The reliability of such measures is also important. Reliability refers to how likely it is that one would find the same measured results if one used the same measure more than once. Though a demo of a robotic performance may work once, the reliability of its performance is another aspect to consider when evaluating its quality. Reproducible results are critical to the evaluation of scientific results.

Reproducible results are also only possible when one clearly and thoroughly explains one's methods such that readers of the publication could easily run the same study on their own to evaluate the original results. By providing more thorough descriptions of the environment, experimental set-up, and participant population, experimental work in robotics research could much more readily build upon existing studies in order to both evaluate, revise, and extend previous theories and experimental designs.

There are many ways to reduce subjectivity in experimental observations, including using double-blind experimental setups in which neither the experimenter nor the participant (if there is one) knows which experimental condition is being run.

III. Variable-Based Research

Scientific experiments are typically informed by theories, hypotheses, or just a hunch that there is some variable (X) that causally influences another variable (Y). These experimental results allow for generating knowledge about how changing X influences Y in a desirable direction. An extremely powerful and applicable explanation of how this notion applies to the study of technologies is Nass's argument for taking a variable-based approach to technologically-oriented research [11]. Nass argues that comparing technology A (e.g., radio) to technology B (e.g., television) does not make sense unless you only care about the difference between the specific technologies. Instead, in order to make more generalizable conclusions, it is even more powerful to figure out exactly which *dimensions* of a technology matter (e.g., size, responsiveness, speed) in terms of influencing the outcomes that one cares about (e.g., performance on task).

The major problem with object-centered approaches to scientific study is that "theories that are specified or operationalized in terms of one technology can never be applied to any other technology, whether extant or potentially available" (p. 49). Object-centered studies confound all of the variables at play within any given technological system (e.g., studying "telephones" rather than "analog, keyboard and receiver input, high fidelity, two-way processing technologies" (p. 62)). In contrast, a variable-based approach to robotic systems can identify the common threads that are woven through sets of systems that systematically influence their quality and performance.

The variable-based approach also goes beyond experimental methods. It also supplies research communities with portable concepts and raises awareness of issues for other contexts. Regardless of whether research takes place in the lab or in the field, the variable-based approach enables "the researcher [to] make conclusions that will apply to a wide range of technologies (and organizations), including technologies (and organizations) that the researcher has never seen and including technologies (and organizations) that do not even exist yet" (pp. 61-62). A variable-based approach offers the ability to form theories around dimensions (e.g., system complexity, interactivity, similarity to humans; task regularity, permanence across space, synchronicity) rather than objects (e.g., PDAs, cell phones, robots).

This variable-based approach also applies to engineering and practical concerns such as organizational decision-making. In terms of technology design, there is a clear role for technological innovation to drive scientific inquiry. "To the extent to which a technical process is an application of scientific knowledge it contributes nothing to science, while empirical technology, which is itself unscientific, may well offer for this very reason important material for scientific study" [12]. In other words, the invention and adoption of new technologies inspires the investigation of new technological dimensions. For example, if there were no radios or telephones, we would not have studied face-to-face communication in terms of dimensions like copresence, visibility, or simultaneity (e.g.,[3]). Similarly, scientific variables can feed back into practical issues such as managerial decision-making: "By considering all the dimensions of a technology, not just the dimensions which one intended to purchase, managers can predict can predict the net effect of a technological change of organizational behavior" ([11], p. 62). Unlike the more typical application of technology-to-science or science-to-technology, this variable-based approach allows for mutual feedback between both domains.

The variable-based approach is one that has been applied to human-computer interaction and human-robot interaction (HRI) to create knowledge that is usable and testable by others in the future, regardless of changes in the specific technologies. To ground these ideas in real experimental robotics work, the following section presents two recently published studies that take a variable-based approach to HRI.

IV. Examples of Variable-Based Research

These two studies present concrete examples of published HRI research. Both studies were designed to be informative for HRI theories and the design of human-robot interactions, thereby benefiting both scientific inquiry and engineering / design work.

In terms of experiment design, both studies employed random assignment of participants to experiment conditions, isolating the causal independent variables of interest. We used between-participant experiment designs, making it extremely difficult for any given participant to guess at what variables were being manipulated in the study. Between-participant experiment designs are also helpful for avoiding issues of reactivity that are sometimes found in within-participant experiment designs.

In terms of the participants, none of the people in the study were labmates or close friends of the experimenters of the study so they were less likely to try to help the experimenters find particular results. A limitation of these studies is that they used a university student population so the results might only generalize to other university student populations. This is explicitly mentioned in the text of the reports and can be addressed by replicating the studies with other populations. We were careful to balance the number of male and female participants in each of the experiment conditions in order to have a more representative sample of participants in relation to the more general population.

Because the effect sizes of these manipulations were unknown, we did not conduct power calculations to determine how many participants to run for each condition. If we were to run follow-up studies, then we could use the effect sizes found in these studies as inputs for the power calculations.

In terms of data analysis and reporting of results, statistically significant differences between experiment conditions were reported. Results approaching significance were noted, but not used to make strong interpretive claims. Because we used standard data analysis techniques (e.g., analysis of variance), the results lend themselves to analytical scrutiny and comparison against related work.

All limitations we could identify were included in the text of these publications so that others can decide how to interpret the findings and how to improve their own future work.

A. Study 1: Self Extension Into Robots

Taking a variable-based approach to HRI, the first study [7] investigated the research questions: How does one's experience with building a robot influence one's interaction with the robot and perceptions of the robot? Also, how does robot form influence a person's experience with the robot and perceptions of the robot? To address these two questions, we designed a 2 (robot form: humanoid vs. car) x 2 (assembler of robot: self vs. other) between-participants experiment (*N*=56) in which people built a robot (either a humanoid or car) and used a robot (either the one they built or one that was supposedly built by someone else) to play a game.

The independent variable of robot form was manipulated by altering the physical form of the Legos Mindstorm robot to be either more human-like or car-like, but use approximately the same parts between the two forms. The manipulation of a self-assembled vs. other-assembled robot was implemented by taking away the robot from the participant after they had assembled it, making some shuffling noises behind a wall, and then returning to the participant with the same robot–either telling the participant that it was the same robot that the participant had assembled or that it was a different robot from the one that the participant had assembled. The cover story for the need to take the robot behind the wall was that all participants had to use comparable robots in the study.

As done in previous work [9], we evaluated people's perceptions of how much the robot's personality overlapped with their own personality as a proxy for understanding how much people experienced a sense of self-extension into the robot. We also measured how attached people felt to the robot, how much in control people felt in the interaction, and how much of a teammate they perceived the robots to be. Each of these attitudinal indices were created through the use of multiple questions. This helped the indices to be more robust to the particular nuances of single-item measures.

Because we were not interested in gender differences, but wanted to be sure to equally address both genders in the general population, we chose to balance gender across each of the conditions. Thus, we accounted for gender balancing in the experiment design, but did not treat gender as an independent variable in the data analyses.

TABLE I
DISTRIBUTION OF PARTICIPANTS (MALE / FEMALE) IN STUDY 1

CONDITION	**humanoid robot**	**car-like robot**	total
self-assembled	7 / 7	7 / 7	14 / 14
other-assembled	7 / 7	7 / 7	14 / 14
total	14 / 14	14 / 14	28 / 28

The procedure for this study involved filling out a pre-questionnaire, building a partially assembled robot (either the humanoid or car-like robot), having the robot taken away behind a wall and returned (as either the "self-assembled" or "other-assembled" robot), play a game with the robot, and filling out a post-questionnaire.

Using a simple analysis of variance (ANOVA), we found that people had more positive experiences and felt a greater sense of self-extension into the robots that they built themselves. Similarly, car-like robots evoked more positive experiences and a greater self-extension into the robots when compared to human-like robots. The humanoid robot was perceived as having greater control over the situation than the car-like robot. Those robots that were self-assembled were also perceived as more like teammates to participants than other-assembled robots.

As much as possible, we held constant all variables that were not of interest. This is consistent with the notion of ceteris paribus in experimental psychology, i.e., manipulating only the variables one is studying and holding all else constant. For example, we made sure that the humanoid and car-like robot forms were similar in size, were made of the same materials (the same Legos Mindstorm kit), and were approximately the same in terms of complexity of assembly.

As with all of our lab's studies, we used standardized measures or slight modifications thereof so that our results would be usable by others and more readily compared against existing literatures. Selecting from standard measures used in previous work of researchers in the field (e.g., from [9]) makes for a good starting point for generating one's set of dependent variable measures.

Implications for both theory and design were discussed. Of primary importance to theory was the support of the hypotheses (1) that less human-like forms of robots would encourage greater self-extension from people and (2) that people self-extend more into robots that they built themselves than robots that were built by others. The implications for design are (1) that less human-like forms of robots and self-assembled robots should be used in situations where self-extension into robots would be positive and (2) that more human-like forms of robots and other-assembled robots should be used in situations where self-extension into robots would be negative.

B. Study 2: Disagreeing Robots

Taking a variable-based approach to HRI, the second study [17] investigated the research questions: How does robot disagreement with a person influence human-robot collaborative task performance? How does the placement of robot voices (speakers) influence the human-robot team performance? To address these questions, we designed a 2 (robot disagreement: none vs. some) x 2 (robot voice location: on robot body vs. in control box) between-participants experiment (*N*=40) in the context of a human-robot desert survival task. This task was used in previous research in human-computer interaction, but was modified for the design of this experiment.

The robot's voice location was manipulated by changing the location of the speaker through which the robot's voice prompts were played. The voice location on the robot body was implemented by placing the speaker on the robot's back. The voice location in the control box was implemented by placing the speaker inside of a hollow box placed on the ground in front of the participant.

Because we did not want the robot's disagreements to be specific to any particular items in the set of survival items, we chose to have the robot disagree on the second, fourth, and fifth item choices of the participant. The robot would describe the item selected by the participant, offer an opinion about its relationship to the alternative selection, and then describe the alternative item. For example, if the participant chose the knife over the pistol, then the disagreeing robot would say, "The knife could be helpful for cutting down stakes to build a solar still or to build shelter. It could also assist in cutting down firewood for a fire. *That is not as good as* the pistol, which could be used for signalling for help. It could also provide an alternative noise source if your voice is weak due to dehydration. Which do you choose?" The agreeing robot would say the same quote, except that, "that is a better choice than," would replace the phrase, "that is not as good as."

We measured decision-making outcomes and attitudes toward the robot in this study. Because the robot only disagreed with participants on the second, fourth, and fifth items, we measured how many times each participant changed their answers on one of those three items to indicate how much the participant's decision was swayed by the opinions of the robot. We used Principle Component Analysis to identify which Likert scale items were inter-correlated to generate indices of perceived robot agreeableness (2 items), sense of similarity with the robot (4 items), and liking of the robot (8 items). Reliability of each index was calculated, using Cronbach's alpha.

Just as in Study 1, Study 2 was not explicitly designed to check for gender differences in how people respond to disagreeing robots, but it was designed to balance the genders of participants across experiment conditions. Thus, we ran five participants of each gender in each of the four conditions. If we had wanted to analyze these data with gender as an independent variable, then we would have run at least ten participants of each gender in each of the four experiment conditions, thus doubling the total number of participants in the study.

TABLE II
DISTRIBUTION OF PARTICIPANTS (MALE / FEMALE) IN STUDY 2

CONDITION	**agreeing robot**	**disagreeing robot**	total
voice on robot	5 / 5	5 / 5	10 / 10
voice in box	5 / 5	5 / 5	10 / 10
total	10 / 10	10 / 10	20 / 20

The procedure for this study involving filled out a pre-questionnaire, reading over the desert survival scenario and writing down one's initial survival item rankings, interacting with the robot to get the robot to fetch the survival items in order of importance, writing down one's final survival item rankings, and filling out a post-questionnaire. The robot behavior was created by a Wizard of Oz methodology [4] in which two people hid in a side room, controlling the robot's motions and triggering pre-recorded voice responses from the robot to the participant.

Once again, using simple ANOVAs, we found that people were more accepting of disagreeing robots if their voices were projected from separate bodies (e.g., control boxes); this was an interaction effect. Participants actually changed their answers more often when the robot disagreed with them than when the robot always agreed with them. They also felt that the robot was more similar to them and more agreeable when it agreed than when it disagreed with them; this was a manipulation check.

We chose a slightly modified version of a standard experimental task (desert survival collaborative decision-making task) that is actually used in real-world team-building settings and has been used across dozens of studies in human-computer interaction in the past. Because the task was modified, we provided the exact script of the robot voice prompts in the text of the publication so that others could use the same task if they wanted to run a replication or follow-up study in the future.

Unlike the first study, we used both behavioral and attitudinal measures so that conclusions could be drawn about both performance on the task and feelings about the experience.

While it is often easier to gauge differences in attitudes about a particular experience with a robot, it is often more important to know if there are differences in behavioral task performance with robots, too.

Implications for both theory and design were discussed. In terms of theory, this study supported the notion that the physical placement of robot voices influences people's behaviors and attitudes about robots; it also supports the notion that robot "opinions," not just facts, can influence decision-making behaviors in people. In terms of design, this study suggests that robots that will have to disagree with people, should have their voices places off of the robot body; it also suggests that judgments made by robots should be carefully designed because people's decisions might actually be swayed by those robots' opinions.

V. Discussion

While these two sample experiments in human-robot interaction provide one example of experimental research in robotics, figuring out how to hold other domains of robotics research to similar standards is still open for discussion.

A. Ongoing Efforts

The current push for benchmarking [5], shared data sets (e.g., computer vision data sets [2]), shared performance metrics (e.g., [16]), and standard testing grounds (e.g., NIST testing sites) are solid steps toward improving the standards of experiments in robotics research. Dillman [5] organizes benchmarks along two dimensions, functional/analytic and component/system types of evaluations, which is a useful taxonomy for organizing robotics system evaluations.

Ongoing work in pushing for more open source software for robotics (e.g., ROS, Player/Stage, YARP) and shared hardware platforms (e.g., Willow Garage's PR2, iRobot's Create, and Activerobot's Pioneer) represent another step toward sharing more research tools to enable generalizeable and repeatable research. In an effort to improve the sharing of such research across labs, Willow Garage is generating and sharing an open source robot operating system (ROS) and will soon be launching a beta program to make a hardware robotic platform available to approximately ten robotics labs so that they can share and test their work more readily.

B. Experiment Designs

Experiment designs become surprisingly complicated when there are too many research questions and/or variables at play. Because it is often expensive to run such experiments, it is tempting to test many hypotheses and variables at once, but this is generally a bad idea because the statistical methods used for analyzing such data become increasingly complicated and difficult to interpret. There are many other aspects of Studies 1 and 2 that we considered, but left out for these reasons. It was safer to save those variables for follow-up experiments than to risk creating needless complexity in these initial study designs.

Experiment design guidelines in the discussion of Studies 1 and 2 include:

- holding all else constant besides the independent variables of interest (ceteris paribus)
- randomly assigning participants to experiment conditions
- balancing the number of female and male participants in each experiment condition
- using standardized and/or previously used tasks
- using standardized and/or previously used measures of dependent variables
- including both task performance and attitudinal measures (in the case of human participants)
- representative sampling of participants from the population of interest (in the case of human participants)

Many other experiment design guidelines not raised earlier include:

- ensuring sufficient data points in each condition, using power calculations (if effect sizes are known) or norms within the research community
- using a variety of instantiations to improve the generalizability of results
- protecting the rights of participants (in the case of human participants) [1]

These guidelines are just a starter set of issues to consider when designing experiments for robotics research.

The emphasis on one-time demonstrations of robot performance (e.g., grand challenges or other competitions) in robotics are one way of comparing the performance of robots, but they do not necessarily prove that one's robotics research is consistently better or worse than another lab's. Furthermore, unless the robots are specifically designed to test the effectiveness of particular aspects of robots (e.g., quadruped vs. biped), then these competitions do not necessarily offer generalizable solutions for future robotics research projects. The generative results of scientific research are contributed by having larger sample sizes of repeatable demonstrations.

Of particular relevance to robotics research is the consideration of how generalizable one's results will be if the study is run with a one-of-a-kind robot hardware. Though it is already a challenge to get one system working for a demo, it is even more difficult to get a system running on a variety of hardware platforms. The benefit of being able to reliably demonstrate one's research contribution on more than one type of robot (e.g., using one's new algorithm, introducing new sensor hardware, etc.) is that the results of the studies are much more generalizable and shareable.

C. Statistical Analyses

One must be very careful when statistically analyzing and interpreting results. A particularly helpful reference book for choosing experiment designs and their statistical analyses is Winer's textbook that occupies the bookshelves of many experimental psychologists and graduate students [18]. This is not an introductory text for beginners in experiment design, but it is a useful handbook for those who already know the basics and want to look into more advanced experiment designs and analysis techniques. More recent developments in dyadic analysis [8] are also useful for understanding pairs of participants rather than just individuals. Many of the statistical analysis techniques from education and agriculture are also helpful for figuring out how to analyze group-level data. If one is at a university, then it is also often helpful to consult with members of the Statistics Department to get advice on both the design of studies and the proper analyses of the data collected from those studies.

One particularly common statistical analysis that must be critically evaluated is statistical modeling. Cognitive modeling in artificial intelligence research makes heavy use of statistical models. Statistical models are useful for being able to show that one's model is better than other models at predicting the variance in a desired outcome (via R-squared comparisons). However, an issue to point out in this domain is that the outcome of such statistical models is highly dependent upon the selection of variables included in the model. Adding a new variable to a model (X2) can even change the direction (+/-) of influence of existing predictors (e.g., X1) upon the desired outcome (Y). This is taught in basic statistics courses and is cause for concern over the casual use of causal models in statistical analyses [14]. Though it is becoming increasingly easy to toss in and swap out predictor variables with the latest statistical analysis software packages, it does not necessarily mean that these are productive work practices. Unless the researcher has very high confidence in his or her selection of predictor variables and relations between them, it does not make sense to use complex causal modeling to analyze one's data.

D. Reporting Methods and Results

There remains an open question of how experiment results should be presented in publications. A goal for such standards of reporting might be enabling others to have sufficient information presented in a transparent enough way such that they could re-run the experiment themselves. If this requires posting data sets online (as done by CMU [2]), then the publications should point to the exact location of the repository of data for others to test their own systems. If this requires particular hardware and/or software, then the means for obtaining those resources should also be made explicit. Including version identifiers of code checked in to subversion repositories would help others to replicate the original authors' work. Although this might seem to run against individual's desires to stay in the lead of a research area, it is actually beneficial to the individual to share one's code and other resources in order to enable others to build upon and extend one's work, contribute to a more open research community, and encourage more researchers to read, use, and cite one's research publications.

One particularly useful example of a research community with strict standards for reporting experiment results is the American Psychological Association (APA). The APA manual is a standard handbook used by anyone who wants to publish in this research community and by all editors of publication venues both inside and outside of the APA [13]. It explicitly states exactly what types of information need to be reported in each section. For example, the methods section of a standard APA submission must include participant populations and demographics, equipment and materials used in the study, and step-by-step procedures for the study. The results section must include informationally adequate statistical results, including specific values that must be reported for each type of statistical analysis such as statistical power, statistical significance, effect size and strength of variable relationships.

Some of the experiment reporting issues included in the discussion of Studies 1 and 2 include:

- explicitly stating research questions and hypotheses
- reporting only statistically significant results, not only descriptive results (e.g., frequencies, means)
- thoroughly reporting on methods and procedures so that readers could run the study on their own
- explicitly discussing limitations of each study

Other experiment reporting issues include:

- thoroughly covering related work to better ground the study in existing literature
- reducing bias in language used for reporting results, e.g., hypotheses are never "proven," only "supported"
- cautiously reporting on statistically non-significant differences; you cannot "prove" that two conditions are the same, using p-values, because non-significant results can also be caused by faulty experiment designs, measures, analyses, etc., not a true lack of difference
- thoroughly and clearly labeling figures and tables
- defining the most important terms, e.g., how was "accuracy" calculated?

Because of the written linguistic format and static imagery used in research publications, it is often difficult to share the richness of the system, the data, and the results, but providing digital repositories for public access on the Internet is one way to share research in other formats, e.g., [2]. Other prestigious scientific publication venues such as *Science* require much more rigorous and thorough paper submissions than the final formats that are ultimately published. The goal is to provide succinct and readable articles for a wide variety of readers to understand, but to also have the rigorous and thorough reporting available elsewhere in case the reader is interested in the richer details of the study. This might be an option to consider for robotics research publications in the future if it would be helpful to reach a broader readership while also addressing the detailed concerns of other researchers within one's own field.

Regardless of how these publications are presented, it is critical to maintain high standards for scientific reporting of robotics research if there are to be more generalizable and re-usable methods and results. By being more forthcoming about one's research process and analysis of data, it becomes possible for the reader to appreciate the thoughtfulness and scientific rigor of one's work. By being vague or secretive about one's research, it is reasonable for the reader to suspect that one has sloppy scientific methods to hide.

E. Experiment Work Practices

While it may seem like experiment design and reporting are straight-forward, what you find in textbooks and academic courses are not exhaustive of the actual experiment work practices one would observe and learn by doing actually conducting research in a lab.

One of the most helpful and yet unreported work practices of experimenters is the use of pilot testing. Because pilot tests are not often thoroughly reported in most research publication venues, it might seem like few researchers use them. However, we often use pilot tests for evaluating our experimental stimuli, whittling down measurement sets to include only those items that are necessary, and getting a rough sense of whether or not we think that the study is worth running. Experiments are very costly in terms of researchers' time so it is worth piloting every part of the study possible before committing to running the full version. As an example, we piloted the robot-building instruction set several times to iteratively improve the simplicity and comparability of the two robot-building sessions. We also pilot tested each of the sets of voice agent recordings in Study 2 before selecting which voice actors we would use in the final study and figuring out which voice prompts had to be re-recorded to improve their audibility. It is helpful to have both very experienced researchers and novices participate in pilot studies because they are likely to comment upon very different aspects of the study design and implementation. If your first few participants have trouble with your experiment design, it is fine to remove their data from the final data set, improve upon your study design, and start fresh with the subsequent participants.

Another work practice in experimental research that is not often discussed is the process of generating hypotheses. This involves identifying the variables that one believes are important and hypothesizing about the causal relationships between those variables. Because so much of the experimental research literature focuses upon the experiments themselves, it is easy to lose sight of why the researchers took it upon themselves to run the studies and test their particular hypotheses in the first place. Some people take a very theory-oriented approach to identifying their variables and causal relationships while others rely upon their observations and intuitions about where the most important variables lie.

Grounded theory [6] is emerging as a useful methodology for research in HCI and HRI (e.g., [10]). Though this particular method comes from sociology rather than from experimental social sciences, it may still prove to be a useful set of rigorous methodologies for robotics research. Grounded theory uses an inductive process to systematically identify conceptual categories and theories about their relationships, using either qualitative or quantitative data. While grounded theory is typically used for understanding social systems of people, its methodology is also potentially useful for identifying important variables and their relationships in robotics. It has been used in human-robot interaction [10] to identify the variables at play in integration of robots into real-world hospital workflows. Grounded theory methods might also be used to identify which variables most effectively improve performance (with or without people) of perception, navigation, manipulation, or other robotics research areas.

VI. Conclusion

In this paper, we began by reviewing the theoretical literature regarding the relationships between science and technology as well as variable-based approaches to research. To ground the discussion in improving experimental methodologies in robotics, we presented two published experiments from the human-robot interaction research literature. We then discussed the broader issues in experimental methodologies, including experiment designs, statistical analyses, reporting methods and results, and other experiment work practices.

Good experimental research requires having the humility to give oneself the chance to be wrong. Testing theories and iterating upon them is the nature of scientific research. Empirically generated research findings are beneficial to scientific knowledge in that they either support one's ideas when the results are consistent with one's predictions, or they guide those ideas in new directions when the results are not consistent with one's predictions. By clearly and thoroughly describing one's experimental research, other scientists can run identical or similar studies to test the validity and generalizability of one's claims.

While generalizability in robotics is difficult to achieve due to the diversity of robotic parts and systems in various research labs, it is not impossible. If one lab does studies on their unique robot, the hypotheses, methods, and measures used by that lab are still usable by others. By testing and replicating others' evaluations of their own robots, the field can move forward to create more general results that apply to multiple types of robots instead of finding particular results that uniquely apply to particular robots.

VII. Acknowledgments

Thanks go to Steffi Paepcke for helping to edit this paper.

References

[1] Collaborative Institutional Training Initiative, http://www.citiprogram.org/

[2] CMU Computer Vision Test Images, http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/cil/ftp/html/v-images.html.

[3] H. H. Clark, *Using Language*, New York: Cambridge University Press, 1996.

[4] N. Dahlback, A. Jonsson, and L. Ahrenberg, "Wizard of Oz Studies: Why and How," *Proceedings of the Conference on Intelligent User Interfaces*, pp. 193-200, 1993.

[5] R. Dillmann, *Benchmarks for Robotics Research*, EURON, 2004.

[6] B. Glaser and A.L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, New Brunswick: Aldine Transaction, 1967.

[7] V. Groom, L. Takayama, and C. Nass, "I Am My Robot: The Impact of Robot-Building and Robot Form on Operators," *Proceedings of the Human-Robot Interaction Conference: HRI 2009*, pp. 31-36.

[8] D.A. Kenny, D.A. Kashy, and W.L. Cook, *"Dyadic Data Analysis"*, New York: Guilford Press, 2006.

[9] T. Kiesler, and S. Kiesler, “My Pet Rock and Me: An Experimental Exploration of the Self Extension Concept,” in *Advances in Consumer Research*, 22 (32), 2004.

[10] B. Mutlu and J. Forlizzi, “Robots in Organizations: The Role of Workflow, Social, and Environmental Factors in Human-Robot Interaction,” *Proceedings of Human-Robot Interaction Conference: HRI 2008*, pp. 287-294.

[11] C. Nass, L. Mason, “On the Study of Technology and Task: A Variable-Based Approach,” in *Organization and Communication Technology*, Ed. J. Fulk and C. Steinfeld, Newbury Park: Sage, 1990, pp. 46-67.

[12] M. Polanyi, *Personal Knowledge: Towards a Post-Critical Philosophy*, Chicago: University of Chicago Press, 1990.

[13] *Publication Manual of the American Psychological Association*, Washington, DC: APA, 2002.

[14] D.R. Rogosa, “Casual Models Do Not Support Scientific Conclusions: A Comment in Support of Freedman,” *Journal of Educational Statistics*, 12, pp. 185-195, 1987.

[15] H. A. Simon, *The Sciences of the Artificial*, Cambridge, MA: MIT Press, 1996.

[16] S. Singh, Special Issue on Quantitative Performance Evaluations of Robotic and Intelligent Systems, *Journal of Field Robotics* 24 (8/9), 2007.

[17] L. Takayama, V. Groom, and C. Nass, “I’m Sorry, Dave: I’m Afraid I Won’t Do That: Social Aspects of Human-Agent Conflict,” *Proceedings of Human Factors in Computing Systems: CHI 2009*, April 2009, pp. 2099-2107.

[18] B.J. Winer, *Statistical Principles in Experimental Design*, New York: McGraw-Hill, 1962.